DOCUMENT RESUME

ED 386 485 TM 024 053

AUTHOR Allen, Nancy L.; And Others
TITLE The Optional Essay Problem and the Hypothesis of Equal Difficulty.
INSTITUTION Educational Testing Service, Princeton, NJ. Program Statistics Research Project.
REPORT NO ETS-RR-93-40; ETS-TR-93-34
PUB DATE Aug 93
NOTE 50p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Difficulty Level; *Equated Scores; *Essays; Essay Tests; Hypothesis Testing; Selection; *Test Items; Test Results
IDENTIFIERS Advanced Placement Examinations (CEEB); *Optional Test Questions

ABSTRACT

A special case of examinee choice, the Optional Essay Problem, is examined from the point of view of test equating. The Optional Essay Problem involves equating essay scores when the examinees are required to select an optional essay topic from a list of topics in addition to taking a mandatory test required of all examinees. The conditions that must be satisfied if the null hypothesis of equal difficulty of the essays holds true are derived. If this hypothesis, called "Livingston's Null Hypothesis," holds true, there is no need to equate the scores. The conditions take the form of inequalities about unobservable quantities that may be displayed graphically. They are illustrated with a real example from the Advanced Placement Examinations. S. A. Livingston's (1988) proposal of adjusting essay scores in the Optional Essay Problem is analyzed and explained from the perspective of test equating, and his proposal is generalized to two new proposals that are explicit about the assumptions they make concerning the unobserved data. These methods are illustrated, and the results for adjusting optional essay scores are used to propose comparable procedures for directly adjusting linear composite scores that include mandatory and optional test scores. Six tables and five figures present analysis data. (Contains 12 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
H. I. BRAUN
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RR-93-40

ED 386 485

The Optional Essay Problem and the Hypothesis of Equal Difficulty

Nancy L. Allen
Paul W. Holland
Dorothy T. Thayer
Educational Testing Service

PROGRAM
STATISTICS
RESEARCH

Technical Report No. 93-34

Educational Testing Service
Princeton, New Jersey 08541

BEST COPY AVAILABLE

TM024063

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants.

Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

The Optional Essay Problem and the Hypothesis of Equal Difficulty

Nancy L. Allen
Paul W. Holland
Dorothy T. Thayer
Educational Testing Service

Program Statistics Research
Technical Report No. 93-34

Research Report No. 93-40

Edcational Testing Service
Princeton, New Jersey 08541

August 1993

Copyright © 1993 by Educational Testing Service. All rights reserved.

Abstract: We examine a special case of examinee choice, the Optional Essay Problem, from the point of view of test equating. The Optional Essay Problem involves equating essay scores when the examinees are required to select an optional essay topic from a list of topics in addition to taking a mandatory test required of all examinees. We derive conditions that must be satisfied if the null hypothesis of 'equal difficulty' of the essays is true. (We call this Livingston's Null Hypothesis.) If this hypothesis holds then there is no need to equate the scores on the optional essays. Our conditions take the form of inequalities about unobservable quantities that may be displayed graphically. We illustrate them using a real example from the Advanced Placement Examinations. Then we analyze Livingston's (1988) proposal for adjusting essay scores in the Optional Essay Problem and explain it from the perspective of test equating. We use our explanation to generalize his proposal to two new proposals that are explicit about the assumptions they make concerning the unobserved data. (We argue that every method for adjusting the essay scores in the Optional Essay Problem must make assumptions about unobserved data.) We illustrate the adjustment methods with an example from the Advanced Placement Examinations. Finally, we use the results for adjusting optional essay scores to propose comparable procedures for directly adjusting linear composite scores that include both a mandatory and an optional test score.

1. INTRODUCTION

Testing programs often have examinations that consist of both mandatory and optional parts. For example, many of the Advanced Placement Examinations have a multiple-choice portion that is required of all examinees and an additional set of essay topics from which each examinee must choose one or more on which to write. The 'optionality' of the essay is only in the topic chosen, not in whether or not to write the essay portion of the exam. One result of these choices is that examinees do not all take the same total test, and, more importantly, their own choices determine important features of which complete test they do take. Allowing examinee choice in such tests is often justified as a way of preventing examinees from having to work on a 'large' test item (like an essay) that they feel is inappropriate for them. Topics can be inappropriate for various reasons, such as a special curriculum used in a course taken by the examinee, or their academic concentration in the humanities versus the sciences.

In the applications that we have in mind, the scores given to the optional portions of the test usually involve some subjective element; human graders evaluate essays or problem solutions and assign scores to them. In this type of grading, it may be difficult to apply exactly common grading standards across different problems or to essays written on different topics. In addition, it may be difficult to construct essay topics or problems that are of equal difficulty. These considerations, in turn, can lead to cases of unfairness to examinees that may undermine the good intentions that justified allowing examinee choice in the first place.

For example, suppose that, in addition to the other mandatory parts of the test, examinees must select one essay topic from a set of five topics. Suppose also that topic 1 is inherently harder than the other topics or that the grading of topic 1 is more stringent than for the other topics. Examinees who had the misfortune to select topic 1 are possibly disadvantaged by their choice. Their scores are lower than they would have been had they chosen another topic. Can we separate the wisdom of an examinee's choice of topics from the effects of unintended differential difficulty or differential grading standards? This is the general problem of interest to us.

Differential severity of essay or problem grading and differences in problem difficulty may not be obvious or intentional and this may make any attempt to regrade all the essays or problems with new grading standards unlikely to produce comparable results. When this occurs, statistical adjustments to the scores (score equating) may be required to achieve a fair test for all examinees.

To give focus to the paper we will consider in detail a special case of examinee choice that we call the Optional Essay Problem. We remark that

we do not limit the applicability of the Optional Essay Problem to cases where the optional tests are actual 'essays'. For example, they also might be math or science problems that the examinees must work out, showing all the steps in their reasoning, which are then graded by one or more problem readers.

THE OPTIONAL ESSAY PROBLEM: Suppose that the complete test is made up of a mandatory part, with raw-score denoted by X, and a single optional 'essay'; and that each examinee must select a topic for the optional essay from a list of K topics. If an examinee chooses topic i, we denote the raw score on essay topic i by Y_i.

The problem, then, is to equate the scores on the optional essay topics so that the examinees are not unfairly disadvantaged if there are differences in difficulty or in the severity of the grading across the topics. The only data available from a single examinee is a pair, (X, Y_i), for some topic i that varies for each examinee.

Our approach is to treat this as a test equating problem with missing data. The missing data are the scores on all the essays that the examinee did not select. If an examinee chooses to write on topic 1, then Y_1 is observed, but Y_2, Y_3, ..., Y_K are all missing for that examinee. The way that we will decide on the need to equate the essay scores is to estimate what the marginal distributions of Y_1, Y_2, ..., and Y_K would be if each examinee had been assigned an essay topic at random, i.e., had exercised no choice of essay topic. If the resulting estimated distributions of some of the Y_i scores are notably different from the others, then equating may be necessary. We will use linear, observed-score equating methods in this paper because they involve only first and second moments of distributions and produce simple linear equating functions (Angoff, 1971; Holland and Rubin, 1982; Petersen, Kolen, and Hoover, 1989). However, the more general observed-score, 'kernel equating' methods described by Holland and Thayer (1989) also fit into the scheme described here. We will summarize some simple facts about linear observed-score equating after we discuss Livingston's Null Hypothesis in section 2.

A basic assumption of our approach is that it makes sense to consider the essay scores for the topics that an examinee did not select as missing data (i.e., data that could have been observed but wasn't). This is a subtle point because it assumes that each examinee could have selected a different essay topic from the one selected. This is an assumption about the strength of the determinants of that choice. There is an implicit 'similarity' between the K 'essay topics' in our approach. This similarity might not be plausible in some instances of examinee choice, and our methods would not necessarily be applicable to such settings. For example, examinees usually select foreign

language Achievement Tests on the basis of the languages that they studied in school (or have other familiarity with). The determinants of an examinee's selection to take a German rather than a French Achievement Test are very strong and it is often implausible to imagine such an examinee selecting a language exam for which they have no preparation. It may not be useful to regard this case of examinee choice as a problem of missing data.

A useful criterion for testing whether or not our approach might apply to a given situation is to ask if it would be appropriate to assign the essay topics at random to the examinees instead of letting them exercise their own choice in the selection. When the choice between topics is relatively hard for examinees to make (i.e., the choices *are not* strongly determined) then random assignment might be appropriate, but when it is easy for examinees to choose between the options (i.e., the choices *are* strongly determined), random assignment is probably not appropriate. Wainer, and Thissen (1993) use 'big choice' and 'little choice' to distinguish between choices that we have described as 'easy' or 'hard.' In practice, when there are several essays topics to choose from, an examinee will find it easy to eliminate some topics from consideration and hard to choose from the rest.

When random assignment of topics is inappropriate, the choices the examinees make become an essential part of the test and this raises important and serious issues of score comparability that are not easily settled either by fiat or by psychometric means, e.g., see Wainer and Thissen (1993).

Finally, we think it is important to remember that the question of whether or not to equate the scores on the optional essays arises only because examinees exercise choice in the selection of topics. If the topics had been assigned at random to the examinees then they would obviously have to be equated and standard random-group methods would be appropriate (Braun and Holland, 1982). One of the interesting things about the Optional Essay Problem is that examinee choice has the dual effect of (a) making the choice of an appropriate equating technique uncertain and (b) calling into question whether or not equating is necessary at all.

The remainder of this paper is organized as follows. Section 2 introduces notation and considers the situation in which it is unnecessary to equate the essay topics, we call this 'Livingston's Null Hypothesis.' In section 3 we derive inequalities that must be satisfied by the data if Livingston's Null Hypothesis is true, and in section 4 we illustrate these ideas with a real data example. In section 5 we show how consideration of the mandatory test score, X, can imply additional inequalities that must be satisfied by the data, and then illustrate these results of using X in section 6. Section 7 addresses what to do if we reject Livingston's Null Hypothesis, and decide to make score adjustments. We first analyze a proposal of Livingston's and then use our analysis to generalize Livingston's approach

to achieve his goals. Section 8 illustrates our proposal and Livingston's proposal on real data. Section 9 generalizes the proposals of section 7 to procedures for adjusting composite scores directly rather than simply adjusting the essay topic scores and then using them interchangeably in a linear composite with X. Finally, section 10 makes a few additional points and summarizes the rest of the paper.

2. LIVINGSTON'S NULL HYPOTHESIS

Livingston (1988) suggests that evidence concerning the null hypothesis of 'equally difficult essay questions' may be used to specify the amount of 'correction' given to the essay scores in the Optional Essay Problem.

> The attempt to produce equally difficult questions may not succeed completely, but in the absence of any statistical information to the contrary it provides a reason for considering raw scores on the alternate questions to be comparable. (Livingston, 1988, page 3)

His position is that if the topic selection, the scoring rubrics, the grading instructions and the training of the essay readers are all carefully implemented then the null hypothesis that the raw scores on the various topics are comparable may be valid and there may be no need to equate or otherwise adjust the essay scores, $Y_1, \ldots, Y_K$. Are there any data routinely collected in the Optional Essay Problem that can shed light on Livingston's Null Hypothesis? In order to answer this question we need to give more precision to its statement, which we shall do in defining LH_0, below.

In observed-score test equating we compute or estimate the marginal distribution of each test score on a common population of examinees. Differences between these distributions are then used to devise adjustments to the scores--e.g., tests with higher mean scores are easier for the population and these scores are, therefore, adjusted downwards while tests with lower mean scores are harder for the population and these scores are adjusted upwards. If there are no differences among the score distributions then no adjustments are necessary since the tests are equally difficult for the population.

We shall interpret Livingston's Null Hypothesis in these terms, but which population shall we use? In the Optional Essay Problem, examinees get to select which topic they will write on, and therefore the sub-population (P_i) choosing essay topic i is non-random and subject to selection. In view of this, we shall use the whole population of examinees (all those taking the mandatory test X) as the population (P) on which to compute the observed-

score equating function. We recognize that in the Optional Essay Problem, self-selection is operating and that this will require us to make assumptions about the distribution of essay scores the examinees would have received if they had chosen to write on a topic different from the ones that they chose. Our approach is to see if these assumptions about selection are compatible with the observed data and Livingston's Null Hypothesis. Furthermore, because we are linearly equating the essay scores we will only concern ourselves with the first and second moments of distributions; we regard two distributions as the same if they have the same mean and variance. This leads to the following precise version of Livingston's Null Hypothesis.

$$LH_0: \quad \mu = \mu_1 = \mu_2 = \ldots = \mu_K, \text{ and}$$

$$\sigma^2 = \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2,$$

where

$$\mu_i = E(Y_i), \text{ and } \sigma_i^2 = Var(Y_i).$$

Thus, μ_i and σ_i^2 are the mean and variance of the essay score Y_i over the whole population (or equivalently, in a large random sample not subject to selection). In our missing data approach to the Optional Essay Problem, we let R_i be the 0/1 indicator for the examinee's choice of essay topic, i.e.,

$R_i = 1$ if the examinee chooses topic i,
$R_i = 0$ otherwise.

In the language of Little and Rubin (1987), the R_i are the missing data indicator variables. The mean and variance of Y_i for the examinees who chose topic i are then defined and denoted by

$$\mu_{i1} = E(Y_i \mid R_i = 1), \text{ and } \sigma_{i1}^2 = Var(Y_i \mid R_i = 1).$$

The quantities, μ_{i1} and σ_{i1}^2, are estimated by the sample mean and variance for the examinees who chose topic i, and in general they are not equal to the population mean and variance, μ_i and σ_i^2 that are referred to in Livingston's Null Hypothesis, LH_0. Along with μ_{i1} and σ_{i1}^2, there are the corresponding quantities for the examinees who did not choose topic i, i.e., for whom $R_i = 0$,

$$\mu_{i0} = E(Y_i \mid R_i = 0) \text{ and } \sigma_{i0}^2 = Var(Y_i \mid R_i = 0).$$

Finally, let

$$p_i = Prob\{R_i = 1\},$$

the probability of selecting topic i, and

$$q_i = 1 - p_i = Prob\{R_i = 0\},$$

the probability of selecting a topic other than topic i.

LINEAR OBSERVED-SCORE EQUATING: One reason for stating LH_0 in terms of the means and variances is that these quantities play a central role in the simplest of all test equating methods--linear observed-score equating. All observed-score equating methods take place on a specific population of examinees (Braun and Holland, 1982). In our problem, this will be the population of all examinees taking the test X, population P. The test scores to be equated, say Y_i and Y_j, however are not observed on this P but on the self-selected sub-populations (P_i and P_j) for which $R_i = 1$ and $R_j = 1$, respectively. Hence, the relevant means and variances needed to equate Y_i to Y_j are not μ_{i1}, μ_{j1}, σ_{i1}^2 and σ_{j1}^2, which are estimated from the self-selected, observed data for Y_i and Y_j. Instead, the relevant means and variances are μ_i, μ_j, σ_i^2 and σ_j^2.

The linear equating function for equating Y_i to Y_j on the population taking X is given by the following well-known linear equating formula:

$$Y_j(y_i) = \mu_j + (\sigma_j / \sigma_i)(y_i - \mu_i), \qquad (1)$$

where y_i denotes a possible Y_i-score and $Y_j(y_i)$ denotes the transformation of this score to the scale of Y_j. This transformation of Y_i-scores will produce scores with the same mean and variance over P as Y_j has. We note that if LH_0 is true then the transformation defined in (1) is simply the identity transformation, indicating that no adjustment is necessary to equate the scores of Y_i to Y_j.

An essential feature of this approach is that it makes explicit the fact that the equating of Y_i to Y_j will involve estimates of μ_i, μ_j, σ_i and σ_j, and that these in turn will involve making assumptions about the missing data. In

our opinion, it is impossible to equate the essays in the Optional Essay Problem without making some sort of assumptions, tacit or explicit, about the missing data. Later we will examine Livingston's (1988) 'ad hoc' procedure for adjusting essay scores in the Optional Essay Problem with this in mind.

3. AN INEQUALITY FOR μ AND σ^2

Theorem 1 summarizes a relationship that must hold between μ, σ^2 and the various quantities defined above when Livingston's Hypothesis holds.

Theorem 1: Under LH_0, we have

$$\sigma^2 = \sigma_{i1}^2 p_i + \sigma_{i0}^2 q_i + (p_i/q_i)\,(\mu_{i1} - \mu)^2 \qquad (2)$$

The proof of theorem 1 is a straight-forward-but-tedious calculation based on computing μ_i and σ_i^2 by conditioning on the two values of R_i, and then replacing μ_i and σ_i^2 by μ and σ^2. Equation (2) expresses the common values, μ and σ^2, assumed in LH_0, in terms of quantities that can be estimated by the observed data, and the unknown variance, σ_{i0}^2, which can not be so estimated.

Our approach to testing Livingston's Hypothesis is to make assumptions about σ_{i0} in terms of its relation to σ_{i1} and to see what implications these assumptions have for μ and σ^2 via equation (2). We define A_i as the ratio

$$A_i = \sigma_{i0}/\sigma_{i1}, \qquad (3)$$

so

$$\sigma_{i0}^2 = (A_i)^2\, \sigma_{i1}^2. \qquad (4)$$

Next we exploit the phenomenon that test score variances are usually quite similar across different sub-populations even though the means of the test scores may vary widely. For example see Table 2 from Holland and Wainer (1990). This can be expressed by inequalities of the form

$$A_L < A_i < A_U, \qquad (5)$$

where A_L and A_U are *a priori* bounds. Examples of plausible values for A_L and A_U might be $A_L = .90$ and $A_U = 1.10$, but other values for these bounds on the A_i may be useful too. We note in passing that a Bayesian analysis in which the A_i are assumed to have a continuous distribution centered on 1 may be developed, but we have not pursued this approach here.

We may combine the inequalities in (5) with the formula given in (2) to obtain conditions that the common mean and variance, μ and σ^2, must satisfy if (5) and LH_0 are true for the specified *a priori* bounds A_L and A_U. These inequalities are given in Theorem 2.

Theorem 2: If LH_0 is true and if the inequalities in (5) are satisfied then μ and σ^2 must satisfy these two inequalities for each essay topic (i = 1 to K),

$$(a)\ \sigma^2 < \sigma_{i1}^2[p_i + (A_U)^2 q_i] + (p_i/q_i)(\mu_{i1} - \mu)^2 \qquad (6)$$

and

$$(b)\ \sigma^2 > \sigma_{i1}^2[p_i + (A_L)^2 q_i] + (p_i/q_i)(\mu_{i1} - \mu)^2. \qquad (7)$$

Statements (6) and (7) define a U-shaped region sandwiched between two parallel parabolas in the (μ, σ^2)-plane, with μ along the horizontal axis and σ^2 along the vertical axis. The two parabolas are defined by the quadratic equations formed from (6) and (7) by replacing the inequalities with equalities. Figure 1 shows the two parallel parabolas for the data for essay topic 2 from Table 1 introduced in section 4, below. In Figure 1, $A_L = .90$ and $A_U = 1.10$.

(Insert Figure 1 about here)

For any essay topic i, these two parabolas share a common vertical line of symmetry, at $\mu = \mu_{i1}$, and differ only in the height of their minima.

The inequalities (6) and (7) together require the possible values for the common mean and variance in LH_0 to lie in this U-shaped region. However, this must hold for each i = 1 to K, so the region of (μ, σ^2)-values that are consistent both with the data and the inequalities in (5) is the intersection of K such U-shaped regions in the (μ, σ^2)-plane, see Figure 2 in section 4, below. Depending on the values of A_L and A_U, this intersection may or may not be empty. If it is empty, then this version of Livingston's Hypothesis is not consistent with the data and the assumptions about the ratios of the

variances in (3) and (5). When this region is not empty it specifies all of the pairs of values of (μ, σ^2) that are consistent with LH_0, the data and our *a priori* assumptions.

4. EXAMPLE 1

Table 1 gives the means, variances and covariances for an example of the Optional Essay Problem, the 1987 administration of the Advanced Placement Examination for European History. In this example, the optional essay topics are topics 2 through 7 while topic 1 is required of all examinees. In the example, topic 1 is ignored.

(Insert Table 1 about here)

Figure 2 shows the 6 pairs of parabolas for $A_L = .90$, and $A_U = 1.10$. Their region of intersection is non-empty and is shaded in Figure 1. Figure 3 shows the 6 pairs of parabolas for $A_L = .95$, and $A_U = 1.05$. The region of intersection for Figure 3 is empty. Thus, in this example, LH_0 is consistent with the data and ratios of standard deviations between 90 and 110 percent, whereas it is not consistent with narrower limits on these ratios, i.e., between 95 and 105 percent.

(Insert Figures 2 and 3 about here)

5. BRINGING X INTO THE PICTURE

Information from the mandatory part of the test may provide information about the relative 'difficulty' of the essay topics. Table 2 displays Livingston's (1988) example of a 'reversal' in the order of the means of the optional essay scores from the order of the means of the mandatory multiple-choice test.

(Insert Table 2 about here)

The point of Livingston's example is that the mean of the multiple-choice score for the group selecting essay 6 (41.5) is the lowest mean of the five groups, but the mean score on essay 6 for those same examinees (7.1) is the highest of the five groups. This is an extreme example of a 'reversal' of the essay and the multiple-choice score means. (There is also an example of a much smaller reversal in Table 1, involving essays 2 and 4.) There is

usually a positive correlation between test scores, so this reversal may suggest that the grading of essay topic 6 is unduly easy relative to the grading of the other essay topics.

Notice that the evidence here involves X, the test score that all examinees have. The inequalities developed in section 3 do not involve X. Our goal now is to derive additional inequalities in the spirit of those of section 3, but which do involve X.

The first step is to reexamine the form of Livingston's Null Hypothesis in LH_0. In the Advanced Placement example of the Optional Essay Problem, the multiple-choice score X is not simply another variable that is obtained from each examinee, but rather it is combined with the essay score to produce a final linear composite raw-score that is then used to form reported scores. Hence, consider the linear composite score

$$S_i = X + w\, Y_i, \tag{8}$$

where $w > 0$ is the relative 'weight' given to the essay part of the composite. The composite score S_i is subscripted with an i because it is based on X and essay i. Note that w does not depend on which essay the examinee selects. Our idea is to replace LH_0 with an equivalent hypothesis about the mean and variance of the composite scores. Again, the distribution of the composites is taken over all of the examinees, not just those selecting essay i. The resulting generalized version of Livingston's Null Hypothesis can be stated initially as:

$$E(S_1) = E(S_2) = \ldots = E(S_K), \text{ and} \tag{9}$$

$$Var(S_1) = Var(S_2) = \ldots = Var(S_K). \tag{10}$$

However, (9) and (10) can be re-expressed in terms of other more basic quantities. First of all, we see that (9) is equivalent to the first part of LH_0 because

$$E(S_i) = \mu_X + w\, \mu_i \tag{11}$$

where μ_X is the mean of X over all of the examinees, and w is non-zero. Secondly, we have

$$Var(S_i) = \sigma_X^2 + w^2\sigma_i^2 + 2\, w\, \sigma_{XY_i} \tag{12}$$

where σ_X^2 is the variance of X over all of the examinees and σ_{XY_i} is the covariance of X and Y_i over all of the examinees. When $i \neq j$,

$$\text{Var}(S_i) = \text{Var}(S_j) \tag{13}$$

if and only if

$$(w/2)[\sigma_i^2 - \sigma_j^2] = \sigma_{XY_j} - \sigma_{XY_i}. \tag{14}$$

We want (14) to hold for all $i \neq j$ and for any value of w that we may choose. The weight, w, is usually determined from considerations external to the question of equating the essay scores and so we require that (14) holds no matter what the choice of w is. This will happen if and only if

$$\sigma_i^2 = \sigma_j^2 \text{ and } \sigma_{XY_i} = \sigma_{XY_j}, \tag{15}$$

for all $i \neq j$.

Combining these results we obtain the Generalized Livingston Hypothesis that is parallel to LH_0.

Theorem 3: If (9) and (10) are to hold for any choice of the weight w then (9) and (10) may be re-expressed as:

$$GLH_0: \quad \mu = \mu_1 = \mu_2 = \ldots = \mu_K, \tag{16}$$

$$\sigma^2 = \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_K^2, \text{ and} \tag{17}$$

$$c = \sigma_{XY_1} = \sigma_{XY_2} = \ldots = \sigma_{XY_K}. \tag{18}$$

The version of the Generalized Livingston Hypothesis expressed in (16)-(18) also can be motivated in other ways, for example, by adding to LH_0 the additional requirement that all the Y_i correlate equally with X over the entire population of examinees.

Next we give the results that parallel Theorems 1 and 2 for the covariance, c, in (18).

Theorem 4: Under GLH_0 we have

$$c = \sigma_{XY_i1}\, p_i + \sigma_{XY_i0}\, q_i + (p_i / q_i)\,(\mu_{Xi1} - \mu_X)\,(\mu_{i1} - \mu) \qquad (19)$$

where

$$\sigma_{XY_i1} = \mathrm{Cov}(X, Y_i \mid R_i = 1), \quad \sigma_{XY_i0} = \mathrm{Cov}(X, Y_i \mid R_i = 0)$$

$$\mu_{Xi1} = E(X \mid R_i = 1), \quad \text{and} \quad \mu_X = E(X).$$

The proof of Theorem 4 parallels that of Theorem 1.

When the correlations between X and Y_i are positive, the natural bounds for points in the (μ, c)-plane are obtained from inequalities for the *ratios* of the correlation of X and Y_i for $R_i = 1$ and $R_i = 0$ as well as the previous bounds assumed on the ratio of the variance of Y_i when $R_i = 1$ and $R_i = 0$. Let B_i be defined by

$$B_i = \rho_{XY_i0}/\rho_{XY_i1} \qquad (20)$$

where ρ_{XY_i0} is the correlation from the covariance $\mathrm{Cov}(X, Y_i \mid R_i = 0)$, etc. Just as for the A_i, it may be plausible to assume that *a priori* bounds for B_i exist, i.e.,

$$B_L < B_i < B_U \qquad (21)$$

for some values of B_L and B_U, such as $B_L = .90$ and $B_U = 1.10$. The next theorem summarizes the resulting two inequalities for c and μ.

Theorem 5: If GLH_0 is true and the inequalities (5) and(21) hold then c and μ satisfy the following two inequalities for all i = 1 to K.

$$(a)\ c < \sigma_{XY_i1}(p_i + q_i\,(\sigma_{Xi0}/\sigma_{Xi1})\,A_U\,B_U) + (p_i / q_i)\,(\mu_{Xi1} - \mu_X)\,(\mu_{i1} - \mu) \qquad (22)$$

and

$$(b)\ c > \sigma_{XY_i1}(p_i + q_i\,(\sigma_{Xi0}/\sigma_{Xi1})\,A_L\,B_L) + (p_i / q_i)\,(\mu_{Xi1} - \mu_X)\,(\mu_{i1} - \mu), \qquad (23)$$

where

$$\sigma_{Xi1}^2 = \mathrm{Var}(X \mid R_i = 1), \text{ and } \quad \sigma_{Xi0}^2 = \mathrm{Var}(X \mid R_i = 0).$$

Thus, σ_{Xi1}^2 is the variance of X in the group who choose essay topic i and σ_{Xi0}^2 is the variance of X for the rest of the examinees.

Inequalities (22) and (23) define a region in the (μ, c)-plane (μ horizontal and c vertical) that is a strip lying between two parallel lines, both with slope

$$-(p_i / q_i)(\mu_{Xi1} - \mu_X). \tag{24}$$

Theorem 5 says that the region of (μ, c)-values that are compatible with GLH_0, the data and the *a priori* bounds A_L, A_U, B_L, B_U is the intersection of these K strips. It is possible, for particular choices of σ_{XY_i1}, σ_{Xi1}, σ_{Xi0}, A_L, A_U, B_L, B_U, μ_{Xi1}, μ_X and μ_{i1} that *no pair* of values of (μ, c) can satisfy all of the inequalities for i = 1, . . . , K. Hence, by bringing the mandatory portion of the test into the picture we double the number of inequalities that must be satisfied. This can make it even harder for GLH_0 to be acceptable in light of the data and our *a priori* assumptions expressed by the bounds A_L, A_U, B_L, B_U.

In Livingston's (1988) analysis he considers the case when the correlations are zero. Our use of ratios of correlations will not work in that case. However, equation (19) is valid even if $\rho_{XY_i1} = 0$, and inequalities similar to (22) and (23) can be developed for this case. The region of possible values for (μ, c) under GLH_0 will then be the intersection of several strips in the (μ, c)-plane that depend on the data, the *a priori* bounds A_L and A_U, and our *a priori* assumptions on the possible sizes of the correlations ρ_{XY_i0}. We do not see how the size of ρ_{XY_i1} can tell us much about the plausibility of either GLH_0 or LH_0, contrary to the position taken by Livingston. In practice we expect ρ_{XY_i1} to be modest and positive and so we will not pursue the case of $\rho_{XY_i1} = 0$ further.

Even though we cannot illustrate the following point with the data in Table 2, inequalities (22) and (23) do allow us to see how 'reversals' like the one in Table 2 might lead to violations of Livingston's Null Hypothesis. The slope (24) of the parallel lines in the (μ, c)-plane specified by (22) and (23) is positive when the mean of the mandatory section scores for examinees who chose essay topic i is less than the overall mean of the scores on the mandatory section, and is negative when it is greater than the overall mean.

In addition, the larger μ_{i1} is the more to the right these lines are shifted. When we have a 'reversal,' as in Table 2, a pair of lines far to the right (large value for μ_{i1}) have large positive slopes rather than the expected negative slopes. This can easily make the intersection region empty so that there are no (μ, c)-values that are compatible with Livingston's Null Hypothesis.

6. EXAMPLE 2

We return to the data in Table 1. The relationship

$$\sigma_X^2 = p_i \sigma_{Xi1}^2 + q_i \sigma_{Xi0}^2 + (p_i / q_i)(\mu_{Xi1} - \mu_X)^2 \qquad (25)$$

allows us to compute σ_{Xi0} from the items in Table 1 for use in (22) and (23). All other values needed are presented in Table 1. Figure 4 shows the resulting pairs of lines for the limits $A_L = .90$, and $A_U = 1.10$, and $B_L = .90$, and $B_U = 1.10$. Figure 5 shows the resulting pairs of lines for the narrower limits $A_L = .95$, and $A_U = 1.05$, and $B_L = .95$, and $B_U = 1.05$.

(Insert Figures 4 and 5 about here)

In both Figures 4 and 5 the intersection of the six regions is non-empty so that there are combinations of μ and c that are compatible with the data and the restrictions (5) and (21).

If we adopt the range $A_L = B_L = .90$, and $A_U = B_U = 1.10$, then Livingston's Hypothesis (expressed either as GLH_0 or LH_0) is compatible with the data in Table 1, and we may conclude that because we cannot reject it there is no need to equate or otherwise adjust the essay scores. For those who think that equating is desirable in this example the onus is to provide evidence that the above bounds on the A_i and B_i are too big. The data in Table 1 can not provide such evidence. Moreover, no data routinely collected in the Optional Essay Problem can provide it. However, through comparisons with other tests and data collected in special experiments (e.g., Wang, Wainer and Thissen, 1993), it may be possible to build up useful prior knowledge for the *a priori* choices of A_L, B_L, A_U and B_U. For example, the 50 observed standard deviations of SAT (V+M)-scores by State given in Table 2 of Holland and Wainer (1990) range from .87 to 1.11 times their mean value of 201. The standard deviations of X and the Y_i in our Table 1 also show a similar small range of values. While these data do not

give direct evidence about the A_i, they do support the intuition that the variation of standard deviations of test scores across populations is often small, and that the choices of $A_L = .90$, and $A_U = 1.10$ do not give an unduly narrow range.

From Figure 2 the resulting range of possible μ -values is roughly between 6 and 8 and from Figure 4 it is roughly between 6 and 9. Thus, $\mu < 6$ is not compatible with these data and with assumptions (5) and (21). We note that the average essay scores for each row in Table 1 all exceed 6 except for essay topic 7 for which the mean is 5.9, slightly below 6. Therefore, the examinees who chose essay 7 are not from the top of the distribution of scores for that essay. According to this analysis, examinees who would have received high scores on essay 7 actually chose other essay topics to write on--and probably got high scores on them. Examinees who did choose topic 7 were from the low end of the score distribution and probably would have received low scores on any essay topic. The question of whether or not the examinees who chose topic 7 would have done better to have chosen a different essay topic can not be answered from the analysis or data presented here.

7. WHAT IF WE REJECT LIVINGSTON'S NULL HYPOTHESIS?

In this section we first discuss Livingston's 'ad hoc' proposal for equating the essays in the Optional Essay Problem, and we then make alternative proposals that are in the same spirit as Livingston's but which are more simply motivated, in our opinion. Our approach involves interpreting various equations given in Livingston (1988) in terms of observed-score test equating. Livingston does not use this interpretation to describe his formulas.

In order to have a simple way of referring to the several populations of examinees that arise in the analysis we remind the reader that P is the entire population of examinees who take X and write on some essay topic from the list of K topics and that P_i is the sub-population of P that writes on topic i.

LIVINGSTON'S 'AD HOC' ADJUSTMENT: Livingston's procedure is fairly complicated, so we break it down into three steps.

Step 1. Equate Y_i to each of the other Y_j and for examinees in P_i obtain the converted value of the observed y_i to the scales of the other Y_j's. Call these converted values $Y_{ij}^*(y_i)$. Livingston uses a special version of equating that we will discuss momentarily.

Step 2. Obtain 'imputed' values, $y_{j,imputed}(y_i)$, for $j \neq i$ for each examinee in P_i. These imputed values are weighted averages of the observed value y_i and its equated value in the Y_j scale, $Y_{ij}^*(y_i)$, of the form:

$$y_{j,\,imputed}(y_i) = (1 - \rho_{XY_j1})\, y_i + \rho_{XY_j1}\, Y_{ij}^*(y_i). \tag{26}$$

Step 3. Compute the adjusted essay score as the simple average of the 1 observed and the K - 1 imputed essay scores for each examinee:

$$y_{adj} = [y_i + \Sigma_{j\neq i}\{y_{j,\,imputed}(y_i)\}]/K. \tag{27}$$

If we define $Y_{ii}^*(y_i)$ as

$$Y_{ii}^*(y_i) = y_i, \tag{28}$$

then we may combine the effect of steps 2 and 3 into a simple expression for y_{adj}, as follows.

Let $\bar{\rho}$ denote the average of all the correlations, ρ_{XY_j1}:

$$\bar{\rho} = [\Sigma_j \rho_{XY_j1}]/K, \tag{29}$$

and let $\bar{Y}_i(y_i)$ be the following weighted average of the converted values:

$$\bar{Y}_i(y_i) = [\Sigma_j \rho_{XY_j1} Y_{ij}^*(y_i)] / [\Sigma_j \rho_{XY_j1}]. \tag{30}$$

We may think of $\bar{Y}_i(y_i)$ as a transformation of y_i into an 'average scale' of the K essay scores determined by the equatings done in step 1 with weights proportional to the correlations, ρ_{XY_j1}. Livingston's final adjusted essay scores, y_{adj}, can be expressed in this notation as

$$y_{adj} = (1 - \bar{\rho})\, y_i + \bar{\rho}\, \bar{Y}_i(y_i). \tag{31}$$

The important feature of Livingston's proposal, in our opinion, is its form expressed in (31) rather than the particulars of its definition. The raw, unadjusted value, y_i, is averaged with the average converted value, $\bar{Y}_i(y_i)$. The weight used in the averaging in (31) reflects Livingston's degree of belief in the relative importance of the converted scores and the original unadjusted score. (Livingston is quite explicit that he regards the evidence

for making an adjustment to the essay score Y_i as the greatest when ρ_{XY_i1} is 1.0, and the least when this correlation is zero.)

We now turn to the equatings referred to in step 1. The method that Livingston proposes for finding the converted scores that we denote by $Y_{ij}^*(y_i)$ may be interpreted as an example of 'equating to a common test,' Angoff (1971), or 'equating through another test,' Braun and Holland (1982) and is often referred to as 'chain equating' by ETS test statisticians (although, in chain equating the operational method used is usually *equipercentile* rather than *linear* equating). The idea is to linearly equate Y_i to X on P_i, then to equate X to Y_j on P_j and finally to compose or 'chain together' these two equatings to get a linear transformation from the Y_i-scale to the X-scale to the Y_j-scale. The first linear equating results in the function:

$$X_i(y_i) = \mu_{Xi1} + (\sigma_{Xi1} / \sigma_{i1})(y_i - \mu_{i1}), \tag{32}$$

The subscript i on $X_i()$ indicates that the equating is on P_i. The second equating results in the function:

$$Y_j(x) = \mu_{j1} + (\sigma_{j1} / \sigma_{Xj1})(x - \mu_{Xj1}), \tag{33}$$

where μ_{Xj1} and σ_{Xj1} are the mean and standard deviation of X for the examinees selecting topic j, defined earlier in Theorems 4 and 5. When the two functions, $Y_j(x)$ and $X_i(y_i)$, are composed or 'chained together' we get

$$\begin{aligned} Y_{ij}^*(y_i) &= Y_j(X_i(y_i)) \\ &= \mu_{j1} + (\sigma_{j1} / \sigma_{Xj1})(\mu_{Xi1} - \mu_{Xj1}) + (\sigma_{Xi1} / \sigma_{Xj1})\,(\sigma_{j1} / \sigma_{i1})(y_i - \mu_{i1}), \end{aligned} \tag{34}$$

which is formula (7) of Livingston (1988), in our notation.

LIVINGSTON'S MISSING DATA ASSUMPTIONS: A key theoretical requirement of test equating is that the resulting equating function should not depend on the population on which it is computed. This gives us a tool for identifying the assumptions about the missing data that Livingston's proposed procedure implicitly makes. Braun and Holland (1982, pg. 37) point out that in order for chain equating to give unbiased results the two equating functions that are chained together, (i.e., (32) and (33)) should not depend on which population is used for the equating. In the present case this means that equating Y_i to X on P_i ought to give the same equating function

as equating Y_i to X on P_j, where, in fact, Y_i is missing data. If we were able to compute the linear equating function of Y_i to X on P_j, the result would be

$$X_{ij}(y_i) = \mu_{Xj1} + (\sigma_{Xj1} / \sigma_{ij1})(y_i - \mu_{ij1}), \tag{35}$$

where

$$\mu_{ij1} = E(Y_i \mid R_j = 1), \text{ and } \sigma_{ij1}^2 = Var(Y_i \mid R_j = 1). \tag{36}$$

The only way two linear functions can be identical is for their slopes and intercepts to be the same. We conclude that the implicit assumptions made in Livingston's proposal are that the missing data, Y_i when $R_j = 1$ and $i \neq j$, satisfies these conditions

$$\sigma_{Xj1} / \sigma_{ij1} = \sigma_{Xi1} / \sigma_{i1}, \text{ or}$$

$$Var(Y_i \mid R_j = 1) = \sigma_{i1}^2(\sigma_{Xj1}^2 / \sigma_{Xi1}^2) \tag{37}$$

and

$$\mu_{Xj1} - (\sigma_{Xj1} / \sigma_{ij1})\, \mu_{ij1} = \mu_{Xi1} - (\sigma_{Xi1} / \sigma_{i1})\, \mu_{i1}, \text{ or}$$

$$E(Y_i \mid R_j = 1) = \mu_{i1} + (\sigma_{i1} / \sigma_{Xi1})(\mu_{Xj1} - \mu_{Xi1}). \tag{38}$$

We may use (37) and (38) to find estimates of μ_i and σ_i^2 that are consistent with Livingston's assumptions about the missing data. They are summarized in Theorem 6.

Theorem 6. If $E(Y_i \mid R_j = 1)$ and $Var(Y_i \mid R_j = 1)$ are given by (38) and (37), respectively, for all i and j, then

$$\mu_i = E(Y_i) = \mu_{i1} + (\sigma_{i1} / \sigma_{Xi1})(\mu_X - \mu_{Xi1}) \tag{39}$$

and

$$\sigma_i^2 = Var(Y_i) = (\sigma_{i1}^2 / \sigma_{Xi1}^2)\, \sigma_X^2. \tag{40}$$

The proof of the theorem follows from multiplying both sides of (37) and (38) by p_j, summing over all j and interpreting the results. These results mean that (a) the mean, μ_i, stands in the same relation to μ_{i1}, as the mean,

μ_X, does to μ_{Xi1}, in terms of the standard deviations, σ_{i1} and σ_{Xi1}, respectively and (b) the ratio σ_i of to σ_{i1} is the same as the ratio of σ_X to σ_{Xi1}.

AN ALTERNATIVE PROCEDURE: Consider an adjusted essay score of the form

$$a_i = (1 - W_i)\, y_i + W_i\, C_i(y_i), \tag{41}$$

where $C_i(y_i)$ is a transformation of y_i to a common scale, and W_i is a weight. We will discuss $C_i()$ and W_i in turn.

In Livingston's procedure $C_i()$ is $\bar{Y}_i()$ defined by (30) and (34). The scale to which this choice of $C_i()$ maps y_i is a weighted average of the scales of all the essays scales using the ρ_{XY_j1} as the weights. The reason for using a weighted average of these scales is to avoid the arbitrary choice of the scale of one of the essay topics as the scale for the other essay topics. The resulting 'average' scale is somewhat unfamiliar. To avoid this we propose using the scale determined by the mean and variance of the total pool of raw essay scores. Let

$$\bar{\mu}_Y = \Sigma_i\, \mu_{i1} p_i \text{ and } \bar{\sigma}_Y^2 = \Sigma_i\, \sigma_{i1}^2\, p_i + \Sigma_i\, (\mu_{i1} - \bar{\mu}_Y)^2\, p_i, \tag{42}$$

then $\bar{\mu}_Y$ and $\bar{\sigma}_Y^2$ are the mean and variance of the entire set of essay scores ignoring that they come from different topics. If we obtain estimates of μ_i and σ_i^2 by making some particular assumptions about the missing data then the transformation

$$C_i(y_i) = \bar{\mu}_Y + (\bar{\sigma}_Y / \sigma_i)(y_i - \mu_i) \tag{43}$$

will map the scale of Y_i to the scale of the raw essay scores with mean $\bar{\mu}_Y$ and variance $\bar{\sigma}_Y^2$.

The transformation given in (43) can be used with any set of assumptions about the missing data that lead to estimates of μ_i and σ_i^2. In particular, using (39) and (40) we can obtain a version of (43) that makes use of Livingston's assumptions about the missing data. There are, however, other alternatives to the chain equating used by Livingston. The most well-known is linear, anchor-test equating in which X is used as the anchor-test rather than as an intermediate test to which Y_i and Y_j are both equated (Angoff, 1971, 1982; Braun and Holland, 1982; Petersen, Kolen and

Hoover, 1989). This method makes explicit assumptions about the missing data that are different from those made by Livingston. These assumptions are related to 'ignorable non-response,' Little and Rubin (1987).

Consider the conditional distribution of Y_i given X and the missing data indicator variable, R_i, i.e.

$$\text{Prob}\{Y_i = y \mid X = x \text{ and } R_i = r\} \tag{44}$$

where r = 0 or 1. If the probabilities in (44) do not depend on r, then the missing data for Y_i is said to be *ignorable given* X. A consequence of ignorability is that the regression function

$$E(Y_i \mid X = x \text{ and } R_i = r) \tag{45}$$

and the variance function

$$\text{Var}(Y_i \mid X = x \text{ and } R_i = r) \tag{46}$$

do not depend on r. If, in addition, we make the further modeling assumptions that the *regression function is linear* and the *variance function is constant* we obtain the two basic assumptions of linear anchor-test equating (also known as Tucker equating):

$$E(Y_i \mid X = x \text{ and } R_i = 0) = E(Y_i \mid X = x \text{ and } R_i = 1)$$

$$= \mu_{i1} + (\sigma_{i1}/\sigma_{Xi1})\rho_{XY_i1}(x - \mu_{Xi1}) \tag{47}$$

and

$$\text{Var}(Y_i \mid X = x \text{ and } R_i = 0) = \text{Var}(Y_i \mid X = x \text{ and } R_i = 1)$$

$$= \sigma_{i1}^2(1 - \rho^2_{XY_i1}). \tag{48}$$

Theorem 7 summarizes the resulting expressions for μ_i and σ_i^2 that follow from (47) and (48).

Theorem 7: If (47) and (48) hold then μ_i and σ_i^2 are given by:

$$\text{(a) } \mu_i = (1 - \rho_{XY_i1})\,\mu_{i1} + \rho_{XY_i1}[\mu_{i1} + (\sigma_{i1}/\sigma_{Xi1})(\mu_X - \mu_{Xi1})] \tag{49}$$

$$\text{(b)}\ \sigma_i^2 = (1 - \rho^2_{XY_i1})\sigma_{i1}^2 + \rho^2_{XY_i1}(\sigma_{i1}^2/\sigma_{Xi1}^2)\sigma_X^2. \qquad (50)$$

We have written (a) and (b) in Theorem 7 in ways that emphasize that under an ignorable (given X) missing data mechanism, μ_i and σ_i^2 are weighted averages (using ρ_{XY_i1} and $\rho^2_{XY_i1}$ as weights) of (a) the estimates of μ_i and σ_i^2 that are implicit in Livingston's procedure (equations 39 and 40), and (b) the raw mean and variance, μ_{i1} and σ_{i1}^2. Livingston's objection to using linear anchor-test equating as the sole basis for adjusting the essay scores is that when $\rho_{XY_i1} = 0$ this approach will result in assuming that

$$\mu_i = \mu_{i1} \text{ and, } \sigma_i^2 = \sigma_{i1}^2 \qquad (51)$$

which will cause a large adjustment to the essay scores when the differences between the μ_{i1} are large; when the $\rho_{XY_i1} = 0$ this is not desirable, in Livingston's opinion.

One way to avoid this is to use an adjusted essay score of the form (41) with

$$W_i = \rho_{XY_i1}, \qquad (52)$$

rather than Livingston's choice of

$$W_i = \bar{\rho}. \qquad (53)$$

Our alternative to Livingston's proposal uses an adjusted essay score of the form (41), using (52) as the weight on the converted score and using (43) to define the converted score. In our view the natural choice for μ_i and σ_i in (43) come from the assumption of ignorable non-response (i.e., Theorem 7) rather than Livingston's proposal to equate Y_i to Y_j through X. However, our approach requires the user to make an explicit assumption about the missing data (i.e., choose an equating method), so using either Theorem 6 or 7 or some other estimates of μ_i and σ_i is compatible with our approach. The result is an adjusted essay score of the form:

$$a_i = (1 - \rho_{XY_i1})\, y_i + \rho_{XY_i1}[\, \bar{\mu}_Y + (\bar{\sigma}_Y/\sigma_i)(y_i - \mu_i)\,], \qquad (54)$$

where μ_i and σ_i^2 are defined in Theorem 7 (if linear anchor-test equating is used to equate Y_i to Y_j) or in Theorem 6 (if chain equating is used to equate Y_i to Y_j).

Our proposal has many of the features of Livingston's proposal given in equation (31). In particular, it tends to dampen the amount of the adjustment that is made to the scores on topic i by the size of the correlation between X and Y_i--the smaller the correlation, the smaller the adjustment. It is also a simple linear function of the essay score alone, i.e., two examinees with identical Y_i-scores but different X-scores will get exactly the same adjusted Y_i-score. It differs from Livingston's in that different assumptions about the missing data, i.e., Theorems 6 or 7, can be used to compute the converted scores, $C_i(y_i)$, from (43). Finally, the scale of the adjusted essay scores in (54) is closely related to the original unadjusted essay scores through $C_i()$.

8. EXAMPLE 3.

Table 3 gives, for the AP European History data in Table 1, the values of μ_i and σ_i under the two different sets of assumptions about the missing data given in Theorems 6 and 7.

(Insert Table 3 about here)

The two sets of assumptions about the missing data give similar estimates of μ_i and σ_i except for μ_3 and μ_7. In addition, all of the estimates of σ_i for the chain-equating assumptions equal or slightly exceed the corresponding σ_i estimates for the anchor-test equating assumptions. The conclusions about the relative difficulty of the essay topics differ somewhat across the two sets of assumptions. For the anchor-test assumptions, topic 3 is the easiest (i.e., least severely graded), topics 2 and 5 are the next easiest, topic 6 the next easiest, and topics 4 and 7 the most difficult (i.e., most severely graded). For the chain-equating assumptions, topics 2, 3, 5, and 7 are the easiest and are about equally difficult, topic 6 is more difficult and topic 4 is the hardest. Both sets of assumptions include topics 4 and 6 among the most severely graded, but they differ substantially on their assessment of the difficulty of topic 7. Topic 3 has the highest estimated mean score under both sets of assumptions, but the anchor-test assumptions assess it as a half a raw-score point easier than the next easiest essay topic. The chain-equating assumptions assess topic 3 as only slightly easier than the other three easy topics.

(Insert Tables 4, 5, and 6 about here)

Tables 4, 5 and 6 give the adjusted scores using Livingston's method and the two methods we have proposed. We used formula (31) for Livingston's method, and formula (54) with μ_i and σ_i given by Theorems 6 and 7 for our two methods. All three methods do not make strong adjustments to the essay scores and if rounded to the nearest integer, Livingston's Ad Hoc procedure and the version of our procedure that uses chain equating make no adjustment at all to the essay scores. The version of our procedure that uses anchor-test equating yields adjusted scores in this example that do not round to the original essay scores in 7 cases. The scores of 1 through 6 for essay topic 3 are adjusted downwards to scores that round down one integer. This is in line with the anchor-test equating estimate of $\mu_3 = 8.0$, making it the easiest of the essay topics. The score 15 for essay topic 6 is adjusted up to a score that rounds up one integer. One might have thought that the anchor-test assumptions would have also resulted in stronger upwards adjustments of the scores for topics 4 and 7. The reason they do not is that the correlations $\rho_{XY_3|1}$ and $\rho_{XY_7|1}$ are smaller than the others--from Y_2 to Y_7 these correlations are: .43, .48, .39, .46, .49, and .37.

Thus, in this example the adjustments are small for all the methods. This is compatible with the results we found earlier regarding the acceptability of Livingston's Null Hypothesis for these data.

9. ADJUSTING COMPOSITE SCORES.

As mentioned earlier, in the applications that we have in mind, such as the Advanced Placement Examinations, the raw scores for the test are weighted composites of score on the mandatory test, X, and the optional essay, Y_i, of the form:

$$S_i = X + wY_i. \tag{55}$$

The X-score will usually be equated by conventional anchor-test methods to older forms of the multiple-choice part, so when we refer to X it is often already converted to an 'X-scale' (although it need not be in particular applications).

The weight , w, is often of the form

$$w = F\,\sigma_X / \bar{\sigma}_Y, \tag{56}$$

so that w is proportional to the ratio of the standard deviation of X to that of the essay scores where no distinction is made between the essay topics. The multiplier F can reflect the relative importance given to the essay score compared to the multiple-choice score X, e.g., F might reflect the amount of time spent on the essay compared to that spent on the multiple-choice part, or possibly the ratio of their reliabilities.

One approach to forming the composite score is simply to replace Y_i by the adjusted score A_i, from (41) or (54), i.e.,

$$S_{i,\,\text{adjusted}} = X + w\,A_i. \tag{57}$$

We might call this the 'plug-in' procedure, for obvious reasons. While probably reasonable in many situations, the plug-in procedure may be objected to on the grounds that the scores that need to be equated are the raw scores that give rise to the reported scores and these are the *raw composite scores* not the *raw essay scores.* An alternative to adjusting the essay scores first and then using the adjusted scores in the composite is to adjust the composite scores directly. In this section we amplify the discussion of section 7 to the case of a composite score.

Usually the weight w can be computed from the data that is on hand prior to any adjustment procedure; in any event, we will assume w is a known value. The composite score, $S_i = X + w\,Y_i$, is very much like the score, Y_i, in that it is only observed for the sub-population of examinees who write on topic i, P_i. We propose an adjusted composite score, s_i^*, of the form:

$$s_i^* = (1 - \rho_{XS_i1})\,s_i + \rho_{XS_i1}[\,\bar{\mu}_S + (\,\bar{\sigma}_S/\,\sigma_{S_i})(\,y_i - \mu_{S_i}\,)\,], \tag{58}$$

where

s_i is the unadjusted composite score from (55),

ρ_{XS_i1} is the correlation of X and S_i on P_i,

$\bar{\mu}_S$ and $\bar{\sigma}_S^2$ are the mean and variance of the composite scores formed without regard to the essay topics involved, and

μ_{S_i} and $\sigma_{S_i}^2$ are the mean and variance of S_i over the whole population of examinees formed by making some assumptions about the

missing data, i.e., the values of S_i for the examinees who did not choose topic i.

As before, we suggest two alternative sets of assumptions that correspond to chain equating and anchor-test equating, respectively. Here are the resulting formulas for μ_{S_i} and $\sigma_{S_i}^2$ for these two sets of missing data assumptions.

Chain equating case:

$$\mu_{S_i} = E(S_i) = \mu_{S_i 1} + (\sigma_{S_i 1} / \sigma_{Xi1})(\mu_X - \mu_{Xi1}) \quad (59)$$

and

$$\sigma_{S_i}^2 = Var(S_i) = (\sigma_{S_i 1}^2 / \sigma_{Xi1}^2)\, \sigma_X^2. \quad (60)$$

Anchor-test equating case:

$$\mu_{S_i} = (1 - \rho_{XS_i1})\, \mu_{S_i 1} + \rho_{XS_i1}[\mu_{S_i 1} + (\sigma_{S_i 1} / \sigma_{Xi1})(\mu_X - \mu_{Xi1})] \quad (61)$$

and

$$\sigma_{S_i}^2 = (1 - \rho^2_{XS_i1})\sigma_{S_i 1}^2 + \rho^2_{XS_i1}(\sigma_{S_i 1}^2/\sigma_{Xi1}^2)\sigma_X^2. \quad (62)$$

In (59) to (62), μ_{S_i1}, $\sigma_{S_i1}^2$, and ρ_{XS_i1} denote the mean, variance and correlation with X of S_i for the examinees who choose topic i. They replace μ_{i1}, σ_{i1}^2, and ρ_{XY_i1} in Theorems 6 and 7, respectively.

The adjusted composite score defined in (58) has the feature that the effect of the equating is dampened by the correlation between X and the composite score, ρ_{XS_i1}. Because S_i contains X, these correlations will usually be much higher than those between X and Y_i and the overall effect will be to put most of the weight on the equated composite scores rather than the unadjusted composite score, s_i.

10. DISCUSSION AND SUMMARY.

Our proposals, e.g., formulas (41), 54), and (58), are 'ad hoc' in the same sense that Livingston's is because there is no real justification for the simple 'weighted average' form of the adjusted scores or for the choice of

weight, $W_i = \rho_{XY_i|1}$. The simple form (41) does have reasonable properties because it dampens the amount of adjustment made to the essay scores, and in the face of the test developers' attempts to make the scores on the optional essays comparable this does seem like a reasonable thing to do. That (41) is a *simple* weighted average is also good because it is easy to understand.

The larger problem is the choice of weight, W_i. It seems to us that W_i should reflect the degree of belief of the user in the assumptions made about the missing data. These assumptions are always untestable when the missing data is really missing, as it is always is in the Optional Essay Problem. Hence, making any particular assumption about the missing data always involves some degree of belief unsupported by the data. Setting $W_i = \rho_{XY_i|1}$ has no basis other than the intuition that the greater $\rho_{XY_i|1}$ is the more plausible it is to believe assumptions underlying the equating. Livingston mentions that any appropriate increasing function of $\rho_{XY_i|1}$ is a possible candidate for the weight. Our analyses do not even suggest that W_i ought to be *any* function of $\rho_{XY_i|1}$, but like Livingston, until there is a better proposal for W_i we think $W_i = \rho_{XY_i|1}$ is a reasonable place to start.

If the correlation between X and Y_i is used as the weight, W_i, in formula (54) it may be useful to consider replacing $\rho_{XY_i|1}$ in (52) by a correlation, ρ_{XY_i}, that has been 'corrected' for 'restriction of range'. This correction can sometimes be substantial. One way to develop a restriction of range correction for $\rho_{XY_i|1}$ is to assume that the missing data for Y_i are ignorable given X, as is done in anchor-test equating discussed in Theorem 7. The result of assuming ignorability given X is the well-known formula (i.e., Pearson, 1903) for the restriction-of-range-adjustment to $\rho_{XY_i|1}$ which is, in our notation,

$$\rho_{XY_i} = \rho_{XY_i|1}(\sigma_X/\sigma_{Xi|1})/[1 - \rho^2_{XY_i|1} + \rho^2_{XY_i|1}(\sigma^2_X/\sigma^2_{Xi|1})]^{1/2}. \quad (63)$$

If (63) is used, then to be consistent the anchor-test equating assumptions for the missing data, i.e., Theorem 7, should be used rather than the chain equating assumptions of Theorem 6. In our example the corrected correlations are: .44, .51, .40, .48, .51, and .37. They are very nearly the same as the uncorrected correlations mentioned at the end of section 8 and the resulting adjusted essay scores are very nearly the same as those in Tables 5 and 6 that use $\rho_{XY_i|1}$ as the weights.

In summary, we have introduced the Optional Essay Problem as a useful paradigm example of examinee choice, and considered the problem of equating the optional essays from two points of view. First, we looked at how we might marshal evidence that the essays don't need to be equated at all (sections 1 to 6). Second, we examined Livingston's proposal for adjusting the essay scores and put it into the context of ordinary test equating (section 7). This has two benefits. First, we can see how to develop several new alternative methods of essay score adjustment that each make different assumptions about the missing data (sections 7 and 8). Second, all of these methods easily generalize to the problem of equating the composite scores of which the optional essays are a part(section 9).

References

Angoff, W. H. (1971) Scales, Norms and Equivalent Scores, in Thorndike, R. L.(ed.) Educational Measurement, 2nd Edition, Washington D. C.: American Council on Education, 508-600.

Angoff, W. H. (1982) Summary and Derivation of Equating Methods Used at ETS, in Holland, P. W. and Rubin, D. B. (eds.) Test Equating, New York : Academic Press, 55-69.

Braun, H.I. and Holland, Paul W. (1982) Observed-Score Test Equating: A Mathematical Analysis of Some ETS Equating Procedures, in Holland, P. W. and Rubin, D. B. (eds.) Test Equating, New York : Academic Press, 9-50.

Holland, P. W. and Rubin, D. B. (eds.) (1982) Test Equating. New York: Academic Press.

Holland P. W. and Thayer, D. T. (1989) The Kernel Method of Equating Score Distributions. ETS RR-89-7.

Holland, P. W. and Wainer, H., (1990) Sources of Uncertainty Often Ignored in Adjusting State Mean SAT Scores for Differential Participation Rates: The Rules of the Game. Applied Measurement in Education, 3(2), 167-184.

Little, R. J. A. and Rubin, D. B. (1987) Statistical Analysis with Missing Data, New York : Wiley.

Livingston, S. A. (1988) Adjusting Scores on Examinations Offering a Choice of Essay Questions. ETS RR-88-64.

Pearson, K. (1903) Mathematical Contributions to the Theory of Evolution-XL. On the Influence of Natural Selection on the Variability and Correlation of Organs. Philosophical Transactions of the Royal Society of London: Series A, 200, 1-66.

Petersen, N. S., Kolen, M. J., and Hoover, H. D. (1989) Scaling, Norming and Equating, in Linn, R. L. (ed.) Educational Measurement, 3rd Edition, New York : American Council on Education and Macmillan, 221-262.

Wainer, H. and Thissen, D. (1993) On Examinee Choice in Educational Testing. Unpublished manuscript under review.

Wang, X. Wainer, H. and Thissen, D. (1993) On the Viability of Some Untestable Assumptions in Equating Exams that Allow Examinee Choice. ETS RR-93-31.

Table 1
Data from the 1987 Advanced Placement European History Exam

Subgroup Selecting Essay	Optional Essay		Multiple Choice			
	μ_{i1}	σ_{i1}	μ_{Xi1}	σ_{Xi1}	σ_{XY_i1}	p_i
2	7.5	2.5	51.9	16.3	17.6	.32
3	8.4	2.3	57.7	15.8	17.4	.13
4	6.5	2.6	53.3	16.6	16.7	.16
5	7.4	2.5	52.0	15.9	18.1	.10
6	6.7	2.4	51.7	16.1	19.0	.11
7	5.9	2.5	42.7	16.9	15.8	.18
Total	7.1	2.6	51.3	16.9	20.2	1.00

Table 2

Data from a hypothetical example showing a reversal, Livingston (1988)

Subgroup Selecting Essay	Optional Essay		Multiple Choice		
	μ_{i1}	σ_{i1}	μ_{Xi1}	σ_{Xi1}	p_i
2	6.4	2.9	45.2	16.7	.06
3	6.9	2.4	47.7	15.6	.65
4	6.6	2.8	47.9	18.1	.02
5	5.9	2.4	48.8	15.6	.18
6	7.1	2.8	41.5	17.4	.09

Table 3
Estimates of μ_i and σ_i using two sets of missing data assumptions.

Subgroup Selecting Essay	Anchor-test equating assumptions		Chain-equating assumptions	
	μ_i	σ_i	μ_i	σ_i
2	7.5	2.5	7.4	2.6
3	8.0	2.3	7.5	2.5
4	6.4	2.6	6.2	2.7
5	7.4	2.5	7.3	2.7
6	6.7	2.4	6.6	2.5
7	6.4	2.5	7.2	2.5

Table 4
Adjusted Essay Scores Using Livingston's 'Ad Hoc' Procedure

Score	Essay Topics					
	2	3	4	5	6	7
1	0.9	0.7	1.4	1.0	1.1	0.9
2	1.9	1.7	2.4	2.0	2.1	1.9
3	2.9	2.7	3.4	3.0	3.1	2.9
4	3.9	3.8	4.4	3.9	4.2	3.9
5	4.9	4.8	5.4	4.9	5.2	4.9
6	5.9	5.8	6.4	5.9	6.2	5.9
7	6.8	6.8	7.4	6.9	7.2	6.9
8	7.8	7.8	8.3	7.9	8.2	8.0
9	8.8	8.8	9.3	8.9	9.2	9.0
10	9.8	9.9	10.3	9.8	10.2	10.0
11	10.8	10.9	11.3	10.8	11.2	11.0
12	11.8	11.9	12.3	11.8	12.2	12.0
13	12.8	12.9	13.3	12.8	13.2	13.0
14	13.8	13.9	14.3	13.8	14.2	14.0
15	14.8	14.9	15.2	14.8	15.2	15.0

Table 5
Adjusted Essay Scores Using (54) and Anchor-test Equating

Score	Essay Topics					
	2	3	4	5	6	7
1	0.7	0.1	1.3	0.8	1.0	1.2
2	1.7	1.2	2.3	1.8	2.0	2.2
3	2.8	2.3	3.3	2.8	3.1	3.2
4	3.8	3.3	4.3	3.8	4.1	4.2
5	4.8	4.4	5.3	4.8	5.1	5.2
6	5.8	5.4	6.3	5.8	6.2	6.3
7	6.8	6.5	7.3	6.9	7.2	7.3
8	7.8	7.6	8.3	7.9	8.3	8.3
9	8.9	8.6	9.3	8.9	9.3	9.3
10	9.9	9.7	10.3	9.9	10.3	10.3
11	10.9	10.8	11.3	10.9	11.4	11.3
12	11.9	11.8	12.3	12.0	12.4	12.4
13	12.9	12.9	13.3	13.0	13.5	13.4
14	13.9	13.9	14.3	14.0	14.5	14.4
15	15.0	15.0	15.3	15.0	15.5	15.4

Table 6
Adjusted Essay Scores Using (54) and Chain Equating
Essay Topics

Score	2	3	4	5	6	7
1	0.9	0.7	1.4	1.0	1.1	0.9
2	1.9	1.7	2.4	2.0	2.2	1.9
3	2.9	2.7	3.4	3.0	3.2	2.9
4	3.9	3.7	4.4	4.0	4.2	3.9
5	4.9	4.8	5.4	5.0	5.2	4.9
6	5.9	5.8	6.4	5.9	6.2	5.9
7	6.9	6.8	7.3	6.9	7.3	7.0
8	7.9	7.8	8.3	7.9	8.3	8.0
9	8.9	8.8	9.3	8.9	9.3	9.0
10	9.9	9.9	10.3	9.9	10.3	10.0
11	10.9	10.9	11.3	10.9	11.3	11.0
12	11.9	11.9	12.3	11.8	12.4	12.0
13	12.9	12.9	13.3	12.8	13.4	13.1
14	13.9	13.9	14.2	13.8	14.4	14.1
15	14.9	15.0	15.2	14.8	15.4	15.1

Figure 1:
Graph of (5) & (6) for AL = 0.90 & AU = 1.10

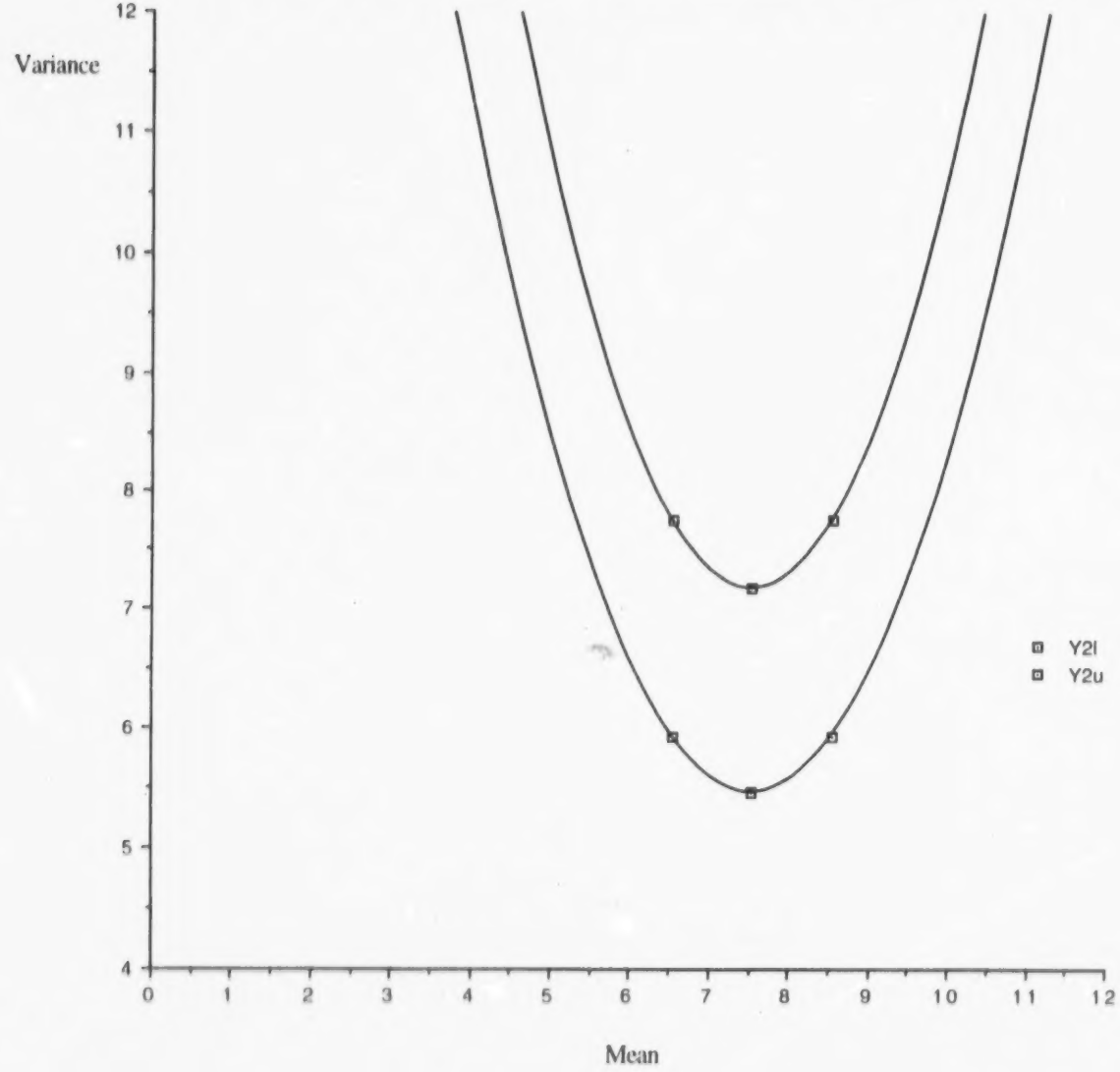

Figure 2:
Graph of (5) & (6) for AL = 0.90 & AU = 1.10

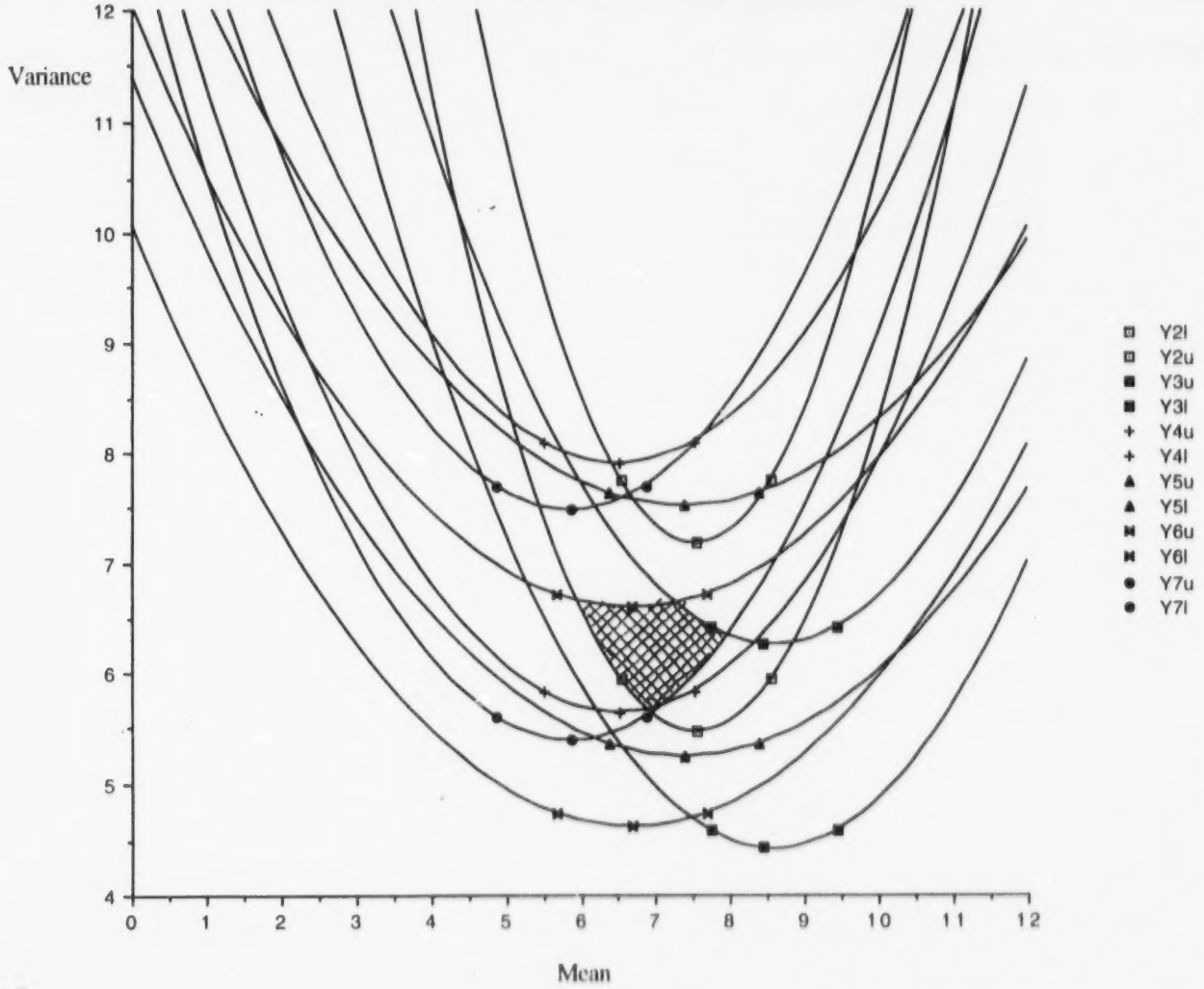

Figure 3:
Graph of (5) & (6) for AL = 0.95 & AU = 1.05

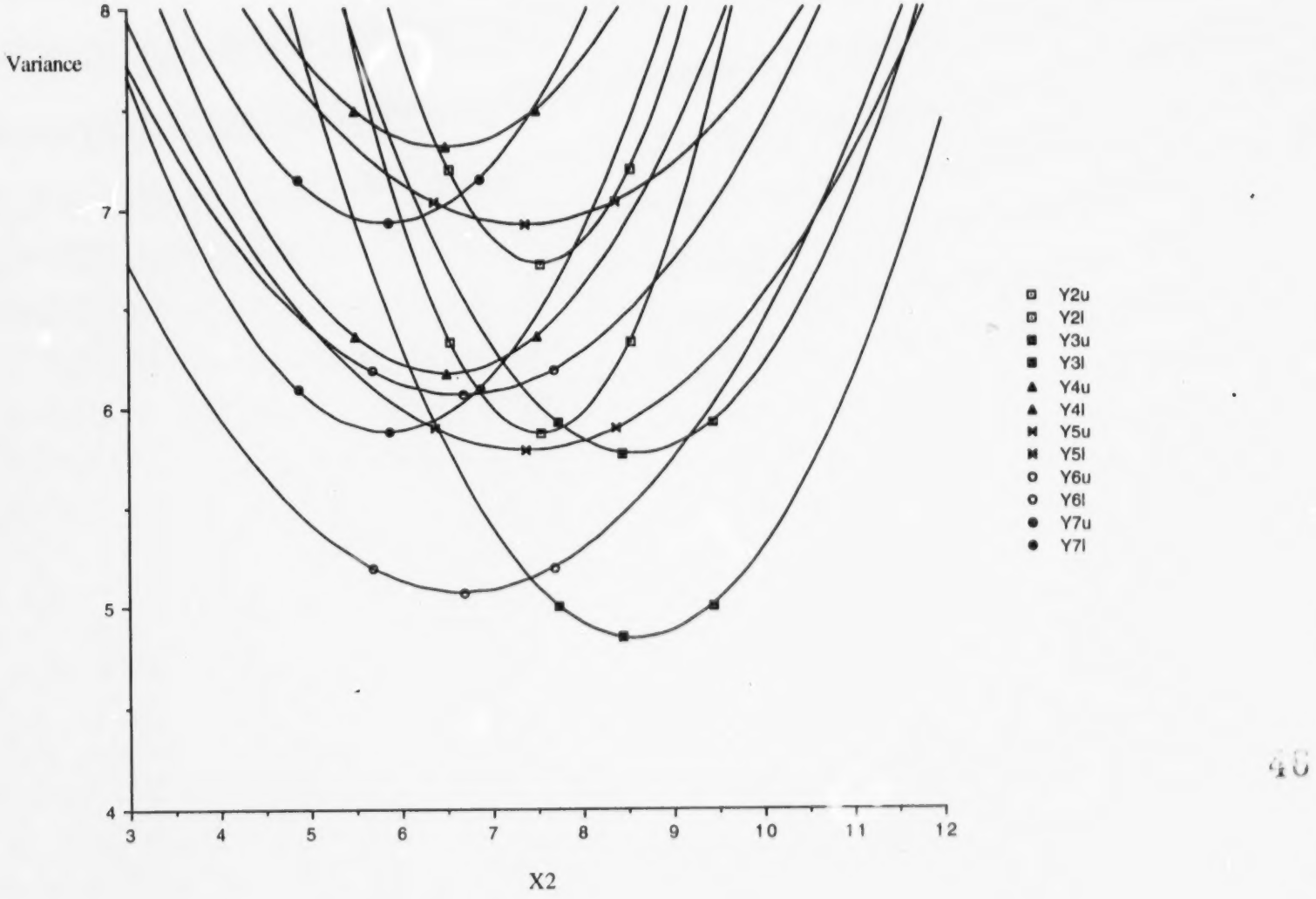

Figure 4:
Graph of (21) & (22) for AL=BL=0.90, AU=BU=1.10

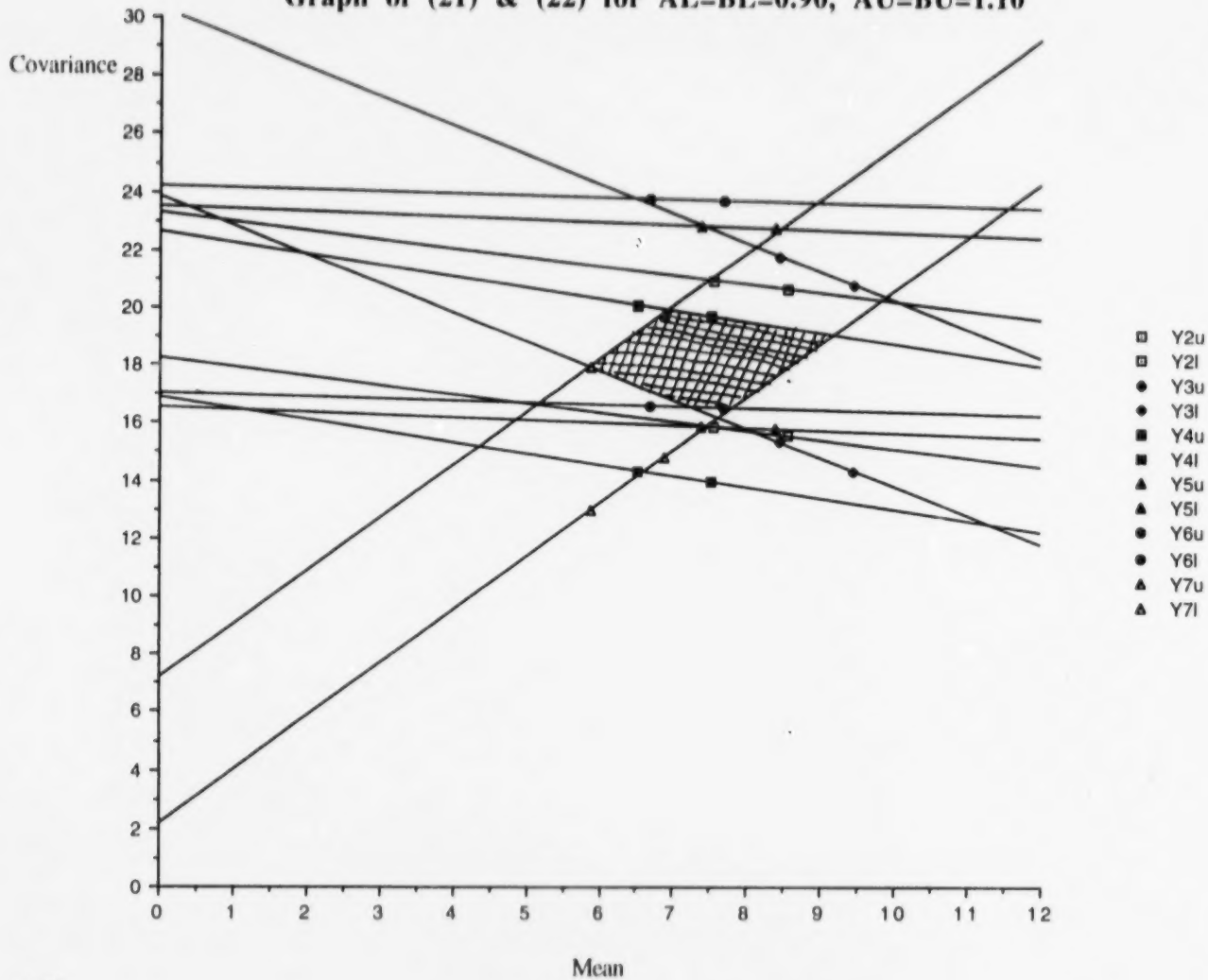

Figure 5:
Graph of (21) & (22) for Al=Bl=0.95, AU=BU=1.05